University of Southern Denmark
IMADA
Arthur Zimek
Jonatan Møller Gøttcke, Jonas Herskind Sejr

DM566/DM868/DM870: Data Mining and Machine Learning

Spring term 2019

Exercise 10: Clustering Algorithms, Density Estimation

Exercise 10-1 Assignments in the EM-Algorithm

Given a data set with 100 points consisting of three Gaussian clusters A, B and C and the point p.

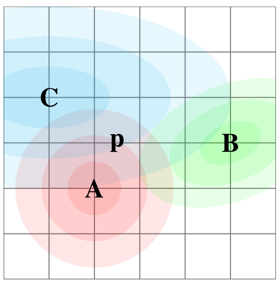
The cluster A contains 30% of all objects and is represented using the mean of all its points $\mu_A = (2, 2)$ and the covariance matrix $\Sigma_A = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$.

The cluster B contains 20% of all objects and is represented using the mean of all its points $\mu_B = (5, 3)$ and the covariance matrix $\Sigma_B = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$.

The cluster C contains 50% of all objects and is represented using the mean of all its points $\mu_C = (1, 4)$ and the covariance matrix $\Sigma_C = \begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix}$.

The point p is given by the coordinates $(2.5, 3.0)$.

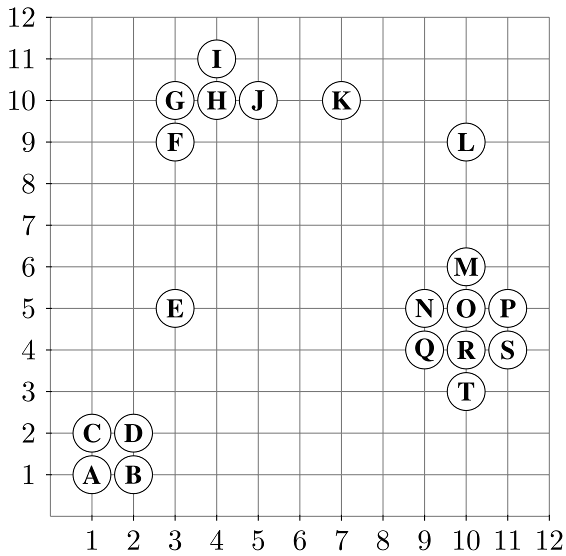
The following sketch is not exact, and only gives a rough idea of the cluster locations:



Compute the three probabilities of p belonging to the clusters A, B, and C.

Exercise 10-2 Density Estimation

Given the following data set:



Estimate the density around each point in the dataset, using the discrete Kernel

$$\hat{f}(x) = \frac{k}{nV_k(x)}$$

based on Manhattan distance (L_1)

(a) with a fixed $k = 2$,

(b) with a fixed $k = 4$,

(c) with a fixed volume based on radius $\varepsilon = 1$,

(d) with a fixed volume based on radius $\varepsilon = 2$.

Explain what your choices are in computing the density estimate regarding

(a) including or excluding the point itself,

(b) ties in the neighborhood.

Note that using the Manhattan distance results in estimators that slightly differ from those discussed in the lecture.

What do you observe?

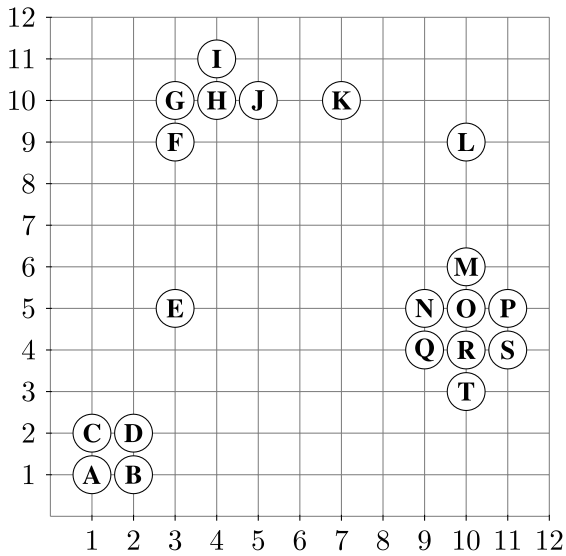
Exercise 10-3 Properties of DBSCAN

Discuss the following questions or statements on DBSCAN:

- For *minPts* $= 2$, what about border points?
- The result of DBSCAN is deterministic for core and noise points, but not for border points.
- A cluster in DBSCAN can contain less than *minPts* objects.
- If the dataset has n objects, DBSCAN computes always exactly n neighborhood range queries.
- On uniformly distributed data, DBSCAN will typically put everything in one cluster or everything in noise. k-means will typically partition the uniformly distributed data in k approximately equal-size partitions.
- What is the relationship of DBSCAN with *minPts* $= 2$ to single-linkage clustering?

Exercise 10-4 **Shared Nearest Neighbors**

Given the following data set:



(a) Compute the pairwise shared-nearest-neighbor-similarities SNN_5 of the objects M, N, O, P, Q, R, S, and T.
Use Manhattan-distance L_1 to obtain the neighbors and neighborhoodsize 5.
The query point is a member of its neighborhood.

$$L_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

(b) Give parameters ε and minpts such that the SNN variant of DBSCAN (Ertöz et al., 2003) identifies the 8 points as "dense" and connects them into a single cluster.